Jurnal Elektronik Ilmu Komputer Udayana
Volume 11, No 3. February 2023

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Pemodelan Topik Artikel Berita Menggunakan *Structural Topic Model* dan *Latent Dirichlet Allocation*

Ayu Kadek Nadya Oktaviana[a1], Ngurah Agus Sanjaya ER[a2], Ida Bagus Made Mahendra[a3], I Gede Santi Astawa[a4], I Gede Arta Wibawa[a5], I Komang Ari Mogi[a6]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Bali, Indonesia
[1]oktaviananadya5@gmail.com
[2]agus_sanjaya@unud.ac.id
[3]ibm.mahendra@unud.ac.id
[4]santi.astawa@unud.ac.id
[5]gede.arta@unud.ac.id
[6]arimogi@unud.ac.id

Abstract

An online news portal is one of the technologies in the form of online media that provides information services in the form of news articles. The number of news articles on online news portals continues to grow over time and more news article data will be available. A large amount of data is a challenge in itself to be processed into a more useful form, namely by conducting topic modeling based on news article data so that the data can be categorized based on the topics discussed in it. Topic modeling groups text data into a specific set of topics based on their similarities. In this study, the dataset used was 44,425 news articles from November 2021 to March 2022 which were taken from the online news portal detik.com. News exploration was carried out by topic modeling using two methods, Latent Dirichlet Allocation (LDA) and Structural Topic Model (STM). The LDA method produces 8 topics derived from the calculation of the highest probabilistic coherence value. The STM method produces 11 topics based on the highest semantic coherence and exclusivity values.

Keywords: *topic modelling, latent dirichlet allocation, structural topic model, news articles, text mining*

1. Pendahuluan

Berita merupakan sarana yang digunakan oleh masyarakat untuk memperoleh informasi mengenai suatu peristiwa atau kejadian faktual secara tertulis. Teknologi yang semakin berkembang menjadikan berita lebih mudah untuk dijangkau oleh masyarakat, salah satunya dengan adanya portal berita *online*. Banyaknya jumlah portal berita *online* yang ada maka artikel berita yang beredar juga semakin banyak. Data berita terus bertambah seiring berjalannya waktu sehingga menjadi tantangan tersendiri untuk mengolah data berita menjadi bentuk yang lebih bermanfaat yaitu dengan menemukan topik tersembunyi dari sekumpulan berita melalui pemodelan topik. Menemukan topik dari sekumpulan berita, dapat menghemat lebih banyak waktu daripada membaca semuanya untuk mengetahui peristiwa yang terjadi dalam jangka waktu tertentu.

Pemodelan topik adalah teknik pembelajaran mesin tanpa pengawasan untuk mengeksplorasi dan mengungkapkan struktur semantik dari sekumpulan dokumen berbentuk teks [1]. Pemodelan topik menggunakan pendekatan *bag-of-words* dimana makna kalimat tidak dievaluasi. Sebaliknya, pemodelan topik mengevaluasi frekuensi kata-kata. Oleh karena itu diasumsikan bahwa kata-kata yang paling sering muncul dalam suatu topik akan menunjukkan tentang topik tersebut [2]. Pemodelan topik telah membuktikan dirinya sebagai alat untuk analisis eksplorasi

sekumpulan dokumen, terutama untuk menemukan topik. Salah satunya adalah penelitian menemukan distribusi topik kemudian klasterisasi dokumen dari cerita berbahasa Bali menggunakan *Latent Dirichlet Allocation* (LDA) [3]. Penelitian lain yang telah dilakukan yaitu menemukan topik dari sekumpulan judul berita menggunakan *Latent Dirichlet Allocation* (LDA) berdasarkan hasil analisis sentimen yang menghasilkan lima topik dari masing-masing sentimen yaitu positif, negatif, dan netral [4]. Selanjutnya penelitian menemukan topik dari portal berita *online* selama masa Pembatasan Sosial Berskala Besar (PSBB) menggunakan metode *Latent Dirichlet Allocation* (LDA) yang menghasilkan empat kelompok besar topik pemberitaan [5].

Penelitian yang disebutkan sebelumnya memiliki kesamaan yaitu menggunakan metode *Latent Dirichlet Allocation* (LDA) untuk menemukan suatu topik dari sekumpulan dokumen. Namun, pemodelan topik menggunakan LDA sering mengabaikan metadata dokumen lain yang bisa mempengaruhi penemuan topik, seperti tanggal publikasi berita, lokasi, penulis, kategori, dan lain-lain. Oleh karena itu pada penelitian ini, penulis melakukan pemodelan topik artikel berita berbahasa Indonesia menggunakan metode lain yaitu *Structural Topic Model* (STM). STM memungkinkan untuk mengeksplorasi korelasi topik dan memahami bagaimana metadata dokumen berhubungan dengan distribusi topik [5]. Penelitian *topic modelling* menggunakan STM telah banyak dilakukan oleh peneliti lain seperti penelitian deteksi topik laten dan tren pada artikel jurnal teknologi pendidikan menggunakan metadata dokumen berupa wilayah negara atau region dan institusi sebagai penentu hasil topik [6]. Penelitian lain menerapkan STM pada 4636 makalah penelitian S2ORC yang berhasil mengungkapkan dua belas topik penelitian, mengetahui korelasi antar topik, dan mengevaluasi topik dari waktu ke waktu [7].

Berdasarkan penelitian-penelitian terdahulu, hasil pada penelitian ini dapat digunakan untuk mengetahui perbandingan topik yang dihasilkan dari metode LDA dan STM dari kumpulan artikel berita berbahasa Indonesia dan mengetahui isu yang terjadi dalam kurun waktu tertentu di masyarakat.

2. Metode Penelitian

2.1. Deskripsi Data

Data yang digunakan pada penelitian ini berasal dari portal berita *online* detik (www.detik.com). Selama lima bulan dari 1 November 2021 sampai 31 Maret 2022, data diambil menggunakan teknik *web scraping* dengan bantuan *framework* Scrapy. Hasil akhir pengumpulan data disimpan dalam format CSV. Tabel 1 menunjukkan deskripsi dari data yang telah dikumpulkan.

Tabel 1. Deskripsi Dataset

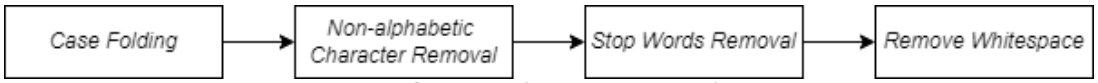
Tanggal	November 2021 – Maret 2022
Sumber	www.detik.com
Tipe Dokumen	Artikel Berita
Fitur	Title, Year, Month, Day, Auhtor, Location, Content
Total	44.425

2.2. *Data Preprocessing*

Preprocessing adalah proses mengolah data yang tidak terstruktur menjadi lebih struktur. Terdiri dari penyetaraan teks, menghilangkan angka, tanda baca, karakter khusus, mengekstrak *stopwords*, dan menghapus spasi kosong. Data teks dibersihkan dengan menghilangkan atau mengubah kata-kata yang tidak bernilai. Semua kata diubah menjadi huruf kecil, dan tanda baca serta spasi dihapus. Karakter khusus dan *stopwords* dihapus karena sering kali tidak berkontribusi pada identifikasi topik. Contoh *stopwords* adalah “dan”, “yang”, dan “ia”. Kata-kata

p-ISSN: 2301-5373
e-ISSN: 2654-5101

ini tidak menambah nilai tentang topik. Gambar 1 menunjukkan tahap *preprocessing* yang dikerjakan pada penelitian ini.



Gambar 1. Tahap *Preprocessing*

2.3. *Document Term Matrix* (DTM)

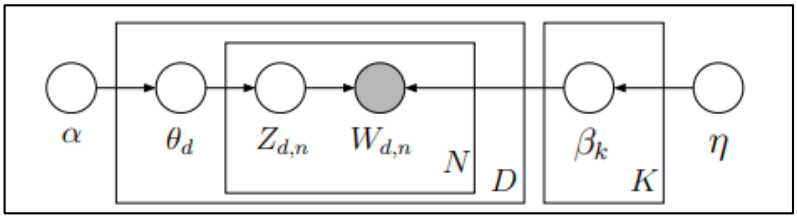
Topic modelling menganggap kata sebagai unit dasar dari sebuah dokumen. Kumpulan kata dalam *corpus* atau sekumpulan dokumen disebut kosa kata (*vocabulary*). *Topic modelling* bergantung pada asumsi *bag-of-words* yang berarti bahwa kata-kata dalam dokumen dapat ditukar sehingga mengabaikan urutan kata dan menghitung frekuensi atau kemunculan kata dari dokumen [8]. Hal ini mengarah pada representasi koleksi dokumen sebagai matriks istilah dokumen atau *document-term-matrix* (DTM) di mana frekuensi kata-kata dalam dokumen dihitung [9]. DTM terdiri dari baris yang mewakili dokumen asli, kolom yang mewakili setiap kata dalam korpus, dan setiap sel berisi frekuensi kemunculan kata tertentu [10].

2.4. *Latent Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation (LDA) mengasumsikan dokumen dan kata-kata sebagai campuran acak atas topik tersembunyi (laten) dimana setiap topik dianggap sebagai distribusi kata yang berasal dari model probabilistik generatif [8]. Model probabilistik generatif bekerja dengan mengamati data, kemudian menghasilkan data yang mirip dengannya untuk memahami data yang diamati. LDA sebagai model probalistik generatif [11]:

(1) Untuk setiap topik,
   (a) Pilih distribusi di atas kata-kata $\beta_k \sim \text{Dir}V(\eta)$.

(2) Untuk setiap dokumen,
   (a) Pilih vektor proporsi topik $\theta_d \sim$ Dir(α).
   (b) Untuk setiap kata dalam dokumen,
      (i) Pilih penugasan topik $Z_{d,n} \sim$ Multinomial(θ_d), $Z_{d,n} \epsilon \{1,..,K\}$.
      (ii) Pilih kata $W_{d,n} \sim$ Multinomial($\beta_{z,d,n}$), $W_{d,n} \epsilon \{1,\ldots,K\}$.

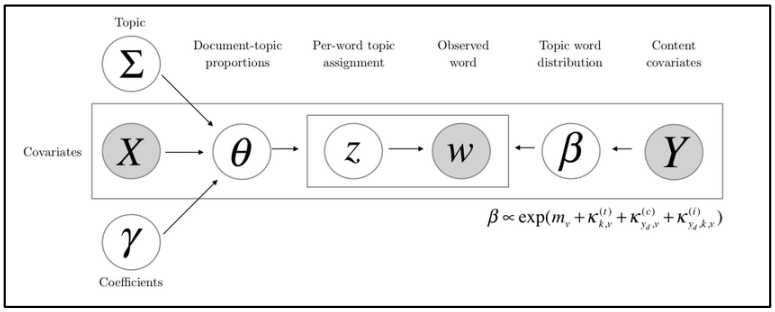


Gambar 2. Model Grafik *Latent Dirichlet Allocation* (LDA)

2.5. *Structural Topic Model* (STM)

Structural Topic Model (STM) merupakan perluasan dari LDA yang mengakomodasi struktur korpus atau metadata dokumen ke dalam model topik [12]. Pendekatan STM dapat digunakan untuk mengeksplorasi korelasi topik dan memahami bagaimana metadata dokumen (kovariat) berhubungan dengan distribusi topik di atas dokumen (prevalensi topik) dan distribusi kata di atas topik (konten topik) [13]. Prevalensi topik adalah tingkat topik yang ada di setiap dokumen. Konten topik adalah frekuensi kata yang digunakan dalam topik.

Pada Gambar 2 yaitu model grafik dari LDA, distribusi topik (θ) bergantung pada parameter *dirichlet* (α) yang menghasilkan distribusi multinomial. Dengan demikian, distribusi topik pada semua dokumen dipengaruhi oleh prior tunggal tersebut. Demikian juga pada distribusi kata (β) terhadap topik juga dikendalikan oleh *dirichlet* sebelumnya (η). Akibatnya, tidak ada kovariat tingkat dokumen yang mempengaruhi distribusi topik (prevalensi topik) dan distribusi kata (konten topik).



Gambar 3. Model Grafik *Structural Topic Modelling* (STM)

Sebaliknya, pada STM prevalensi topik dapat dipengaruhi oleh satu atau lebih kovariat tingkat dokumen (X), yang menghasilkan distribusi unik atas topik untuk setiap dokumen (Gambar 3). Hal ini dilakukan untuk menerapkan *logistic normal linier prior* [$\theta \sim \text{LogisticNormal}(X)$] daripada menerapkan *dirichlet prior* sebelumnya [$\theta \sim \text{Dir}(\alpha)$] ke prevalensi topik dimana *logistic normal linier prior* adalah fungsi dari tingkat dokumen kovariat seperti tanggal terbit artikel, penulis, dan lokasi. Selain itu, STM memungkinkan penggunaan kata dalam topik (konten topik) lebih bervariasi menurut kovariat kategorikal (Y) menggunakan fungsi *multinomial logit* [14].

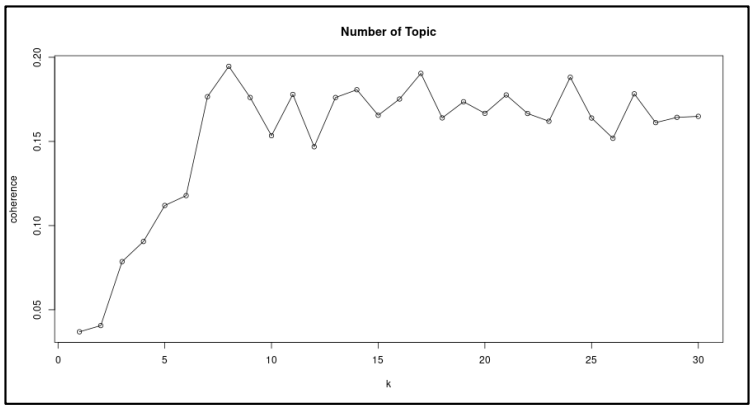
3. Hasil dan Pembahasan

3.1. Menentukan Jumlah Topik Optimal

Salah satu tantangan utama pemodelan topik adalah mengidentifikasi jumlah topik optimal yang laten dalam korpus. Jumlah topik ini digunakan sebagai masukan dalam menjalankan *topic modelling.*

a. *Latent Dirichlet Allocation* (LDA)

Jumlah topik yang optimal untuk metode LDA ditentukan berdasarkan nilai *probabilistic coherence* dengan menguji 30 kandidat topik. Berdasarkan Tabel 2 dan Gambar 4 di bawah ini, jumlah topik yang optimal adalah 8 topik dengan nilai *coherence* terbesar yaitu 0.19387522.



Gambar 4. Hasil Evaluasi Jumlah Topik LDA

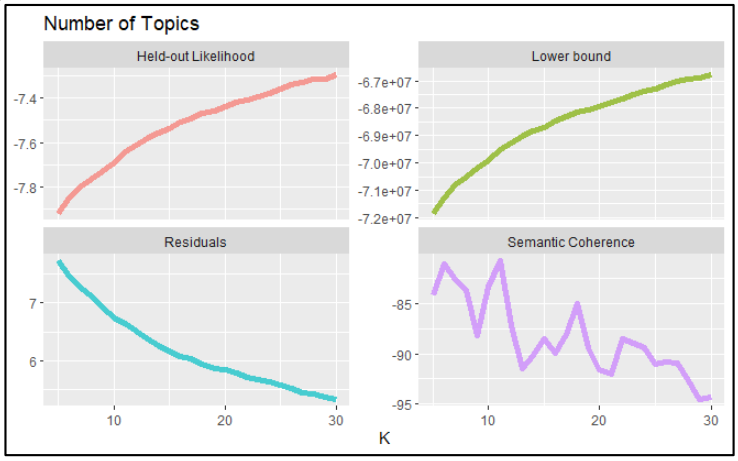
b. *Structural Topic Model* (STM)

Jumlah topik dalam metode STM ditentukan berdasarkan dua kriteria yaitu *semantic coherence* dan *exclusivity* yang keduanya harus setinggi mungkin. Pemilihan kedua kriteria ini dikarenakan hasil dari perhitungan *semantic coherence* akan menghasilkan beberapa topik di mana kata-kata yang sangat umum mendominasi. Oleh karena itu, perlu mempertimbangkan eksklusivitas dari topik untuk memberikan perbandingan. *Semantic coherence* menunjukkan seberapa konsisten suatu topik atau seberapa sering istilah yang menggambarkan suatu topik terjadi bersamaan dan muncul bersama dalam sebuah dokumen. *Exclusivity* adalah seberapa eksklusif topik yang muncul dengan probabilitas

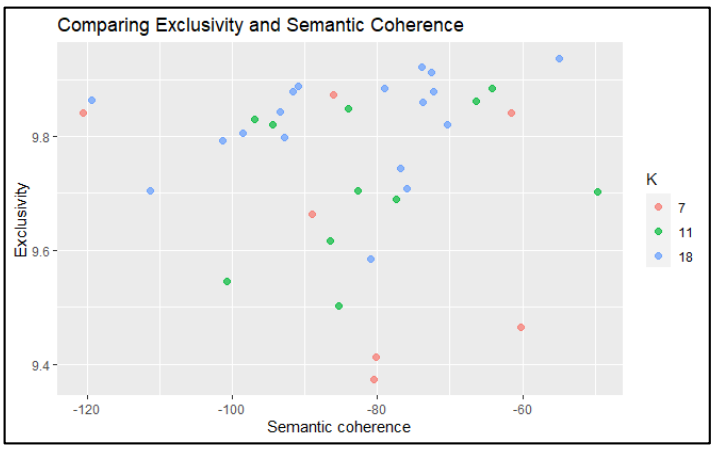
p-ISSN: 2301-5373
e-ISSN: 2654-5101

tinggi untuk suatu topik atau seberapa berbeda topik satu sama lain, dan topik tersebut menggambarkan hal yang berbeda. Semakin tinggi nilai koherensi semantik dan eksklusivitas kata-kata dalam suatu topik, semakin "baik" model topik tersebut.

Gambar 5 menunjukkan hasil evaluasi model STM. Dari Gambar 5 terlihat bahwa jumlah topik tertinggi berdasarkan *semantic coherence* pada kisaran topik 7, 11, dan 18. Untuk memilih topik yang optimal diantara ketiga jumlah topik ini, pertimbangkan juga nilai eksklusivitas dari ketiga topik tersbut yang ditunjukkan pada Gambar 6. Dari Gambar 6, jumlah topik yang optimal adalahh 11 topik dengan nilai koherensi dan nilai eksklusivitas yang tinggi.



Gambar 5. Hasil Evaluasi Jumlah Topik STM



Gambar 6. Perbandingan Nilai *Exclusivity* dan *Semantic Coherence*

3.2. Identifikasi Topik

a. *Latent Dirichlet Allocation* (LDA)

Tabel 2 menunjukkan hasil pemodelan topik LDA berdasarkan jumlah topik yaitu K=8. Nilai koherensi untuk kedelapan topik tersebut memiliki nilai yang beragam. Dari topik dengan nilai tertinggi yang mudah diinterpretasikan, hingga topik dengan nilai koherensi terendah yang sulit untuk ditafsirkan secara sekilas.

Topik 1 membahas mengenai kasus korupsi yang diidentifkasi dengan kata-kata seperti "kpk", "uang", dan "pasal". Hal ini menunjukkan bahwa Topik 1 membahas kasus korupsi. Topik 2 berkaitan dengan kegiatan olahraga, khususnya sepak bola, yang digambarkan dengan kata-kata "gol", "pemain", "laga", dan lainnya. Topik 3 merupakan salah satu topik yang memiliki nilai koheren paling rendah diantara topik lainnya. Dari kata-kata yang dihasilkan, secara ambigu bahwa topik ini membahas mengenai peristiwa yang terjadi di jalanan diidentifikasi dari kata "jalan", "mobil", dan kejadian di daerah yang digambarkan oleh kata "kecamatan", "desa", "air". Topik 4 adalah topik dengan nilai koherensi paling

rendah sehingga kata-kata yang dihasilkan lebih sulit untuk diinterpretasikan dibandingkan topik lainnya. Topik 4 mencakup berita yang terjadi di ibukota Jakarta termasuk pembangunan, program, pendidikan, dan lain-lain. Topik 5 membahas isu COVID19 dimulai dengan vaksinasi, varian terbaru Corona yaitu Omicron, dan menjelaskan cara penanganan PPKM karena terdapat kata "level" menandakan masih ada kabar kenaikan atau penurunan level PPKM di berbagai daerah. Topik 6 berkaitan dengan harga minyak goreng di Pasaran. Topik ini memberitahu bahwa dalam kurun waktu lima bulan yaitu November 2021 sampai Maret 2022 terdapat berita mengenai permasalahan harga minyak goreng di Indonesia. Topik 7 membahas mengnai pemerintah Indonesia yang diidentifikasi dengan kata-kata seperti "presiden", "dpr", dan "partai". Topik 8 terkait dengan negara Rusia dan Ukraina dimana diketahui bahwa kedua negara ini memiliki konflik yang dapat diindentifikasi dari kata "militer", "pesawat" dan "wilayah".

Tabel 2. Hasil *Topic Modelling* dengan LDA

Topik	Coherence	Prevalence	Top Terms
1	0.191	13.243	pasal, pidana, kpk, uang, bukti, laporan, jaksa, dugaan, perkara, jakarta
2	0.401	10.512	pemain, tim, gol, laga, pertandingan, menit, poin, bola, liga, kali
3	0.082	21.569	jalan, rumah, lokasi, desa, mobil, kecamatan, kabupaten, kota, air, kejadian
4	0.061	10.993	kota, jakarta, kerja, dki, daerah, pembangunan, program, pendidikan, kepala, nomor
5	0.111	9.787	covid, kesehatan, kota, vaksinasi, kabupaten, vaksin, corona, omicron, varian, level
6	0.169	13.585	harga, minyak, pt, perusahaan, pasar, juta, goreng, ekonomi, produk, bank
7	0.218	10.521	ketua, jokowi, dpr, presiden, anggota, tni, partai, ri, komisi, politik
8	0.318	9.789	rusia, ukraina, as, dunia, militer, china, presiden, pesawat, wilayah, amerika

b. *Structural Topic Model* (STM)

Hasil pemodelan topik menggunakan STM ditunjukkan oleh Tabel 3. Kata-kata yang tercantum pada Tabel 3 berasal dari keluaran STM dengan jumlah topik K = 11. Tabel tersebut memiliki empat jenis pembobotan kata untuk setiap topik yaitu *Highest Prob*, FREX, *Lift*, dan *Score* [15]. *Highest Prob* adalah kata-kata dengan probabilitas tertinggi atau kata yang paling umum untuk topik tertentu, namun tidak eksklusif dalam arti kata-kata tersebut dapat dikaitkan dengan sejumlah topik. FREX (*frequency-exclusivity*) memberi bobot pada kata-kata berdasarkan seberapa eksklusif kata-kata tersebut terhadap suatu topik pada frekuensi keseluruhannya. *Lift* membobot kata-kata dengan membaginya dengan frekuensinya di topik lain sehingga memberi bobot lebih pada kata-kata yang lebih jarang di topik lain. *Score* mirip dengan *lift*, membagi frekuensi log kata dalam topik dengan frekuensi log kata dalam topik lain.

Tabel 3. Hasil *Topic Modelling* dengan STM

Topic	Proportion	Top Terms			
		Highest Prob	FREX	Lift	Score
1	0.098	video, akun, media, agama, keluarga, suara, foto, sosial, acara, maaf	mui, wayang, yahya, nu, pbnu, gus, yaqut, lagu, ustaz, munarman	bimasislam, dikaruniai, pdri, tarian, mempelai, aurel, slavina, antiteror, najah, akhyar	nu, mui, munarman, wayang, gus, agama, pbnu, yaqut, masjid, arteria
2	0.058	harga, minyak, goreng, juta,	minyak, goreng, liter, harga, bbm,	bersubsidi, kedelai, menstabilkan,	minyak, goreng, harga, liter,

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Topic	Proportion	Top Terms			
		Highest Prob	**FREX**	**Lift**	**Score**
		pasar, ribu, barang, uang, pedagang, kenaikan	kemasan, kg, kedelai, tempe, kripto	bakunya, rti, dow, nurwan, jualnya, pengecer, crude	pasar, kedelai, tempe, kripto, pedagang, curah
3	0.082	covid, kesehatan, vaksinasi, vaksin, corona, omicron, varian, level, kota, persen	vaksinasi, vaksin, omicron, ppkm, dosis, booster, pcr, corona, pasien, prokes	antigen, spesimen, faskes, adisasmito, immunity, herd, penyebarannya, terjangkit, penularannya, eswatini	omicron, vaksin, vaksinasi, corona, covid, varian, ppkm, pasien, dosis, booster
4	0.080	jokowi, presiden, ketua, dpr, partai, ri, anggota, pendidikan, kota, komisi	ikn, pemilu, puan, demokrat, seleksi, fadli, ruu, beasiswa, jokowi, gerindra	pileg, bawaslu, imin, parpol, maharani, ahy, pilpres, fadli, civitas, kelulusan	partai, jokowi, pemilu, ikn, dpr, politik, puan, snmptn, ruu, gerindra
5	0.083	pasal, pidana, kpk, jaksa, hakim, perkara, uang, tindak, dugaan, terdakwa	kpk, jaksa, terdakwa, pn, persidangan, pidana, suap, kejaksaan, korupsi, penjara	dakwaannya, disidangkan, ezer, stepanus, narkotika, narapidana, novia, bnn, lpsk, terpidana	kpk, terdakwa, pidana, jaksa, pasal, penyidik, penjara, korupsi, perkara, hakim
6	0.127	jalan, desa, kota, kabupaten, air, lokasi, hujan, rumah, kecamatan, banjir	sampah, bpbd, longsor, banjir, pohon, hujan, ruas, petir, jembatan, genangan	jpo, bandang, bpbd, majenang, ambarawa, menggenang, bpjt, lengkong, candipuro, lor	banjir, hujan, desa, tol, kecamatan, sungai, kabupaten, longsor, bmkg, bpbd
7	0.056	rusia, ukraina, as, wilayah, militer, china, barat, pesawat, presiden, kapal	rusia, ukraina, putin, gempa, invasi, rudal, kapal, pesawat, biden, israel	kharkiv, belarusia, associated, pemberontakan, artileri, istanbul, kyiv, yahudi, kuleba, pentagon	rusia, ukraina, putin, gempa, invasi, rudal, militer, as, nato, kiev
8	0.068	jakarta, dki, tni, nomor, gubernur, kerja, anies, jenderal, peraturan, sesuai	buruh, ump, upah, mk, formula, tni, bpjs, cipta, ketenagakerjaan, minimum	kartiko, ciptaker, kis, menaker, pangkostrad, feo, jkp, kspi, onlyfans, permenaker	anies, ump, tni, dki, upah, buruh, mk, formula, uu, dudung
9	0.129	mobil, kejadian, rumah, motor, pria, diduga, luka, jalan, ditemukan, lokasi	tkp, mayat, pengemudi, reskrim, polsek, sopir, pengeroyokan, anggiat, handi, kompol	sabetan, diautopsi, inafis, dagu, rakitan, spion, bejatnya, memaki, tusuk, jatanras	mobil, luka, motor, tewas, zulpan, polsek, reskrim, kombes, pengeroyokan, arteria
10	0.109	pemain, tim, gol, laga, pertandingan, menit, poin, bola, liga, kali	pemain, gol, laga, pertandingan, bola, liga, mandalika, juara, motogp, pelatih	newcastle, persik, napoli, ducati, pss, suporter, arsenal, league, juventus, persebaya	gol, laga, pertandingan, pemain, liga, gawang, piala, pelatih, motogp, kemenangan
11	0.110	perusahaan, pt, ekonomi, program, bank, digital, produk, kerja, triliun, keuangan	fitur, nasabah, ekosistem, triliun, digital, baterai, kredit, umkm, oppo, asuransi	multiplier, ekuitas, consumer, baterainya, warjiyo, payment, property, lifestyle, app, emission	triliun, umkm, fitur, digital, produk, listrik, bank, nasabah, saham, pt

Dari Tabel 3, Topik 1 dapat diinterpretasikan dengan lebih mudah diinterpretasikan berdasarkan kata-kata FREX: ‘mui’, ‘wayang’, ‘nu’, ‘pbnu’, ‘gus’, ‘lagu’, ‘ustaz’. Jika dihubungkan dengan kata-kata Highest Prob Topik 1, topik ini terkait dengan agama Islam yang diidentifikasi dari kata ‘mui’, ‘nu’.

Topik 2 memiliki kata-kata FREX dan Highest Prob yang hampir mirip. Oleh karena itu, topik ini mengangkat isu harga minyak goreng. Topik ini juga muncul pada hasil pemodelan topik dengan LDA. Topik 2 memiliki kata ‘harga’, ‘minyak’, ‘kenaikan’. Dari hal ini, dapat disimpulkan bahwa terjadi kejadian kenaikan harga minyak goreng di Indonesia.

Topik 3 yang dihasilkan dengan metode STM juga berhasil muncul di Topik 5 dari hasil pemodelan topik dengan LDA, yaitu membahas mengenai isu kesehatan terutama COVID-19. Keunggulan STM sendiri adalah menghasilkan berbagai kata yang tidak ditemukan pada hasil LDA, seperti kata-kata ‘antigen’, ‘spesimen’, ‘faskes’, ‘immunity’, ‘herd’, ‘penyebarannya’, ‘terjangkit’, dan ‘penularannya’.

Topik 4 berisi pembahasan yang sama dengan Topik 7 dari hasil pemodelan topik dengan LDA, yaitu membahas mengenai Pemerintah Indonesia. Berita ini bisa berupa berita tentang pemerintahan, pemilu, hingga ibu kota baru, yang ditunjukkan dengan kata ‘ikn’.

Topik 5 menghasilkan topik yang mirip dengan Topik 1 pada hasil LDA yaitu Korupsi. Secara sekilas dengan meilhat kata-kata FREX, Lift, hingga Score dapat dipahami bahwa topik ini adalah berita tentang kasus korupsi di Indonesia.

Topik 6 membahas mengenai bencana alam yang terjadi di Indonesia. Diidentifikasi dengan kata-kata seperti ‘banjir’, ‘bpbd’, dan ‘longsor’. Hasil pemodelan topik menggunakan LDA, topik yang berhubungan dengan bencana alam secara ambigu dihasilkan dari satu topik yaitu Topik 3. Topik 3 tersebut membahas dua kejadian dalam satu topik, tetapi pemodelan topik dengan STM, berhasil memisahkan dua kejadian tersebut menjadi dua topik sehingga topik lebih mudah dimengerti.

Topik 7 memiliki kesamaan dengan Topik 8 dari hasil pemodelan topik dengan LDA yaitu isu yang terjadi antara negara Rusia dan Ukraina. Hasil FREX mencantumkan kata ‘invansi’ dan ‘rudal’, yang menunjukkan bahwa kedua negara ini sedang berperang.

Topik 8 membahas mengenai berita yang terjadi di ibu kota Jakarta, ditandai dengan kata ‘jakarta’, ‘anies’ dan lain-lain. Topik ini juga berhasil dihasilkan dari pemodelan topik menggunakan LDA pada topik 4.

Topik 9 adalah pengembangan dari Topik 3 hasil pemodelan topik dengan LDA yaitu kejadian yang terjadi di jalan. Dari hasil Topik 9 terlihat bahwa sering terdapat berita mengenai kejadian di jalan yang diidentifikasi dari kata ‘jalan’, ‘mobil’, ‘motor’, ‘luka’ yang menunjukkan bahwa kecelakaan lalu lintas merupakan hal yang umum diberitakan dalam kurun waktu 5 bulan.

Topik 10 memiliki kesamaan dengan Topik 2 dari hasil pemodelan topik dengan LDA. Topik ini menjelaskan mengenai olahraga khususnya sepak bola yang dijelaskan dengan kata-kata seperti ‘bola’, ‘pemain’, ‘gol’. Kata eksklusif untuk topik ini juga menjelaskan olahraga lain seperti ‘motogp’ dan ‘mandalika’. Seperti yang diketahui bahwa pertandingan MotoGP diadakan pada bulan Maret 2022 di sirkuit Mandalika Lombok.

Topik 11 membahas mengenai ekonomi jika dilihat dari keempat bobot yaitu Highest Prob, FREX, Lift, dan Score. Topik ini diwakili oleh kata-kata seperti ‘bank’, ‘ekonomi’, ‘perusahaan’, ‘digital’, dan ‘keuangan’.

Dari 11 Topik yang dihasilkan dari metode STM, jika dilihat secara sekilas STM menghasilkan topik yang mirip dengan LDA. Topik 3 dan 4 yang dihasilkan dari LDA memiliki nilai koherensi rendah sehingga sulit untuk diinterpretasi secara sekilas. Topik-topik tersebut juga muncul dari hasil STM dan hasil topik dari STM ini lebih mudah dipahami dengan melihat kata-kata teratas dan eksklusif (FREX) kata dari topik tersebut.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

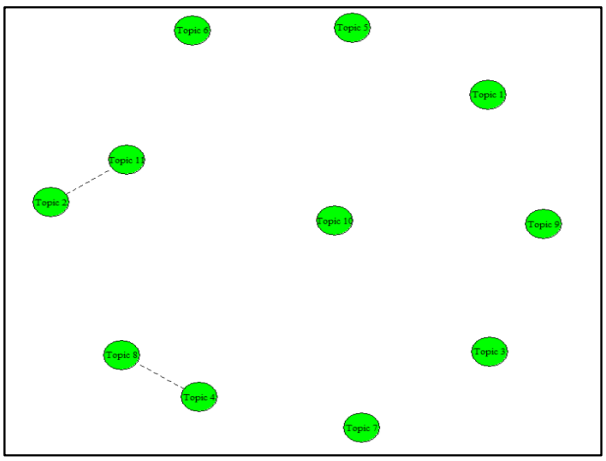
Seperti Topik 3 dari hasil LDA yang membahas mengenai peristiwa yang terjadi di jalanan diidentifikasi dari kata “jalan”, “mobil”, dan kejadian di daerah yang digambarkan oleh kata “kecamatan”, “desa”, “air”. Pada STM, topik ini berhasil dipisah menjadi dua topik yaitu Topik 6 mengenai bencana alam dan 9 mengenai kecelakaan lalu lintas. Kemudian Topik 4 hasil LDA mengenai berita yang terjadi di ibukota Jakarta, topik ini muncul sebagai Topik 8 dari hasil STM.

3.3. Korelasi Antar Topik

Topik yang berkorelasi dapat diidentifikasi menggunakan STM dengan menganalisis seberapa sering topik muncul bersamaan dalam dokumen yang sama. Topik model adalah model campuran yang setiap topiknya memiliki “nilai”, sehingga topik mungkin saja tidak terkait dan berkorelasi. Korelasi positif menunjukkan bahwa dokumen yang memiliki konten tentang topik A cenderung juga menyajikan beberapa konten topik B. Korelasi negatif menunjukkan bahwa dokumen yang menyajikan topik A secara sistematis tidak menyajikan topik B. Semakin banyak topik yang ekslusif atau tidak tumpang tindih, maka korelasi akan menuju nilai -1.

Untuk memudahkan melihat korelasi antar topik dapat dengan melihat Gambar 7. Pada Gambar 7 tersebut, menunjukkan dua node yang terhubung artinya topik tersebut memiliki nilai korelasi yang tinggi. Terdapat empat topik yang saling terhubung yaitu korelasi antara Topik 11 dan Topik 2, dan korelasi antara Topik 8 dan Topik 4.

Topik 11 dan Topik 2 memiliki nilai korelasi 0,06. Jika dilihat dari kata-kata yang dihasilkan dari kedua topik ini memiliki kesamaan mengenai masalah ekonomi dimana Topik 2 membahas mengenai permasalahan harga minyak goreng dan Topik 11 yang membahas mengenai ekonomi digital. Topik 8 dan 4 memiliki nilai korelasi 0,04 dan memiliki kesamaan mengenai Pemerintahan. Topik 8 berkaitan dengan Pemerintahan Indonesia yang diidentifikasi dari kata-kata seperti ‘dpr’ dan ‘jokowi’. Sedangkan Topik 4 membahas mengenai berita yang terjadi di Ibukota Jakarta yang ditunjukkan oleh kata-kata seperti ‘anies’ dan ‘jakarta’. Kedua topik ini rata-rata terjadi di Jakarta.



Gambar 7. Korelasi Antar Topik

4. Kesimpulan

Kesimpulan penelitian ini didasarkan pada dataset yang digunakan yaitu 44.425 artikel berita dari bulan November 2021 hingga Maret 2022 diambil dari portal berita online detik.com dengan melakukan *topic modelling* menggunakan Latent Dirichlet Allocation (LDA) dan Structural Topic Model (STM). Metode LDA menghasilkan 8 Topik yang berasal dari perhitungan *probabilistic coherence* nilai tertinggi. Metode STM menghasilkan 11 Topik berdasarkan nilai *semantic coherence* dan *exclusivity* tertinggi. Dari kedua hasil topik masing-masing metode, LDA memiliki 2 dari 8 Topik yang tidak mudah diinterpretasikan. Sedangkan STM berhasil memudahkan interpretasi topik tersebut dengan melihat eksklusivitas topiknya. Pada metode STM, terdapat 2 pasang topik yang saling berkorelasi yaitu Topik 11 dan Topik 2 memiliki kesamaan mengenai masalah ekonomi. Topik 8 dan 4 memiliki kesamaan mengenai pemerintahan.

Referensi

[1] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimed. Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, 2019, doi: 10.1007/s11042-018-6894-4.

[2] C. B. Asmussen and C. Møller, "Smart literature review: a practical topic modelling approach to exploratory literature review," *J. Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0255-7.

[3] N. A. Sanjaya ER, "Implementasi Latent Dirichlet Allocation (LDA) untuk Klasterisasi Cerita Berbahasa Bali," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 127, 2021, doi: 10.25126/jtiik.0813556.

[4] C. Naury, D. H. Fudholi, and A. F. Hidayatullah, "Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM," *J. Media Inform. Budidarma*, vol. 5, no. 1, p. 24, 2021, doi: 10.30865/mib.v5i1.2556.

[5] W. Wahyudin, "APLIKASI TOPIC MODELING PADA PEMBERITAAN PORTAL BERITA ONLINE SELAMA MASA PSBB PERTAMA," *Semin. Nas. Off. Stat.*, vol. 2020, no. 1, pp. 309–318, 2021, doi: 10.34123/semnasoffstat.v2020i1.579.

[6] X. Chen, D. Zou, G. Cheng, and H. Xie, "Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education," *Comput. Educ.*, vol. 151, no. September 2019, p. 103855, 2020, doi: 10.1016/j.compedu.2020.103855.

[7] R. B. Mishra and H. Jiang, "Management and organizational research: Structural topic modeling for a better understanding of theory application," *Sustain.*, vol. 14, no. 1, 2022, doi: 10.3390/su14010159.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 4–5, pp. 993–1022, 2003, doi: 10.1016/b978-0-12-411519-4.00006-9.

[9] M. Ponweiser, "Latent Dirichlet Allocation in R," no. May, pp. 2–21, 2012, [Online]. Available: http://epub.wu.ac.at/3558/.

[10] A. K. Johnson, R. Bhaumik, D. Nandi, A. Roy, and S. D. Mehta, "' Is this Herpes or Syphilis ?': Latent Dirichlet Allocation Analysis of Sexually Transmitted Disease-Related Reddit Posts During the COVID-19 Pandemic," 2022, doi: https://doi.org/10.1101/2022.02.13.22270890.

[11] D. M. Blei and J. D. Lafferty, "TOPIC MODELS," 2009. doi: 10.1007/978-981-16-0100-2_7.

[12] M. E. Roberts *et al.*, "Structural topic models for open-ended survey responses," *Am. J. Pol. Sci.*, vol. 58, no. 4, pp. 1064–1082, 2014, doi: 10.1111/ajps.12103.

[13] J. Bohr and R. E. Dunlap, "Key Topics in environmental sociology, 1990–2014: results from a computational text analysis," *Environ. Sociol.*, vol. 4, no. 2, pp. 181–195, 2018, doi: 10.1080/23251042.2017.1393863.

[14] E. (Olivia) Park, B. (Kevin) Chae, and J. Kwon, "The structural topic model for online review analysis: Comparison between green and non-green restaurants," *J. Hosp. Tour. Technol.*, vol. 11, no. 1, pp. 1–17, 2020, doi: 10.1108/JHTT-08-2017-0075.

[15] M. E. Roberts, B. M. Stewart, and D. Tingley, "Stm: An R package for structural topic models," *J. Stat. Softw.*, vol. 91, no. 2, 2019, doi: 10.18637/jss.v091.i02.